



# The importance of increasing population diversity in genetic studies of type 2 diabetes and related glycaemic traits

Inês Barroso<sup>1</sup>

Received: 31 December 2020 / Accepted: 7 July 2021  
© The Author(s) 2021

## Abstract

Type 2 diabetes has a global prevalence, with epidemiological data suggesting that some populations have a higher risk of developing this disease. However, to date, most genetic studies of type 2 diabetes and related glycaemic traits have been performed in individuals of European ancestry. The same is true for most other complex diseases, largely due to use of ‘convenience samples’. Rapid genotyping of large population cohorts and case–control studies from existing collections was performed when the genome-wide association study (GWAS) ‘revolution’ began, back in 2005. Although global representation has increased in the intervening 15 years, further expansion and inclusion of diverse populations in genetic and genomic studies is still needed. In this review, I discuss the progress made in incorporating multi-ancestry participants in genetic analyses of type 2 diabetes and related glycaemic traits, and associated opportunities and challenges. I also discuss how increased representation of global diversity in genetic and genomic studies is required to fulfil the promise of precision medicine for all.

**Keywords** Genetics · Genome-wide association study (GWAS) · Meta-analysis · Multi-ancestry · Polygenic risk score (PRS) · Precision medicine · Review

## Abbreviations

GRS	Genetic risk score
GWAS	Genome-wide association studies
LD	Linkage disequilibrium
PRS	Polygenic risk score

## Introduction

Type 2 diabetes is a multifactorial disease where a combination of genes, lifestyle and environment contribute to disease predisposition worldwide. The IDF projected that by 2045, 700 million people worldwide would have diabetes, with low- and middle-income countries accounting for the largest increases in prevalence [1].

Since 2005, with the advent of genome-wide association studies (GWAS), the number of genetic loci known to influence type 2 diabetes risk and/or related quantitative glycaemic measures

(e.g. glucose, insulin, HbA<sub>1c</sub> levels) has surged. To date, more than 270 loci (with >400 signals) associated with type 2 diabetes risk and/or glycaemic traits have been identified mostly through meta-analysis of existing GWAS [2–6]. Despite this success, most type 2 diabetes GWAS do not represent the diversity of affected individuals as they have focused on individuals of European ancestry [7, 8] and, more recently, East Asian ancestry [2, 3]. This means we are still missing important aetiological factors that may differ among diverse global populations and consequently we may be increasing health disparities. In healthcare, sociocultural self-reported ethnicity or ‘race’ are often used as proxies for genetic ancestry. This is particularly problematic, as these proxies are confounded by socioeconomic status and cultural and lifestyle factors, and do not consider the genetic heterogeneity between individuals of the same self-reported ethnicity [9]. Different self-reported ethnic groups overlap genetically and two individuals of the same self-reported ethnicity may be genetically more ‘distant’ from one another than two individuals that each identify with a different ethnic group. This review aims to highlight the opportunities and challenges of including datasets from a broad range of population ancestries in genetic studies of type 2 diabetes and related traits. It also discusses how increasing diversity in genetic studies may impact on precision medicine in type 2 diabetes.

✉ Inês Barroso  
ines.barroso@exeter.ac.uk

<sup>1</sup> Exeter Centre of Excellence for Diabetes research (EXCEED),  
University of Exeter Medical School, Exeter, UK

## Glossary of terms

**Allelic heterogeneity** The process through which different mutations at the same locus can lead to the same phenotype. It also describes the situation in which different variants can underlie the effect of a given locus on a phenotype, in particular when referring to the possibility of the existence of causal variants at the same locus that differ between ancestries

**Clumping** Grouping together variants that represent the same ‘effect’ into a locus associated with a phenotype. Clumping is normally done by linkage disequilibrium (LD) and/or distance to the lead variant. Distance-based clumping normally groups together all variants within a particular region (e.g. a 500 Kb–1 Mb region), centred on the lead variant

**Fine-mapping** The process through which researchers attempt to resolve association signals into a smaller set of variants most likely to be causal

**Genetic risk score (GRS)** A genetic score that is built on the basis of variants associated with a phenotype at genome-wide significant levels. Each variant is weighted by its effect size on disease risk

**Genome-wide association study (GWAS)** An approach that tests, in a single experiment, thousands to millions of genetic variants for association with a phenotype, which can be a disease or continuous trait

**Lead variant** The variant in a locus associated with the phenotype with the most significant  $p$  value

**Linkage disequilibrium (LD)** A measure of the correlation of genetic variants with each other in the genome. Due to differences in population history and demography, LD varies between populations

**Meta-analysis** A statistical approach that allows data from different studies to be combined to increase power. This approach is commonly used to combine GWAS datasets from multiple cohorts to increase power for locus discovery. Combining data across different studies normally relies on the availability of genotype data for the same variants in all studies so as to maximise the data without too many missing values. Usually, this requires imputation of variants not directly assayed in the genotyping array used

***Imputation** A statistical approach that allows the estimation of genotypes at variants not directly assayed. This requires the use of an imputation reference panel that contains genotype data across millions of variants and allows the correlation structure, or LD structure, in the relevant reference population to be used to estimate the missing genotypes in the desired population*

**Partitioned genetic scores** Genetic scores that are built on a subset of variants in a GRS that have been subgrouped based on their inferred effect on different pathways linked to a disease. Thus, the GRS is partitioned into subgroups

**Polygenic risk score (PRS)** Similar to a genetic score but built instead on the basis of thousands to millions of variants in the genome associated with a disease or trait at a more liberal  $p$  value threshold than genome-wide significance

**Winner’s curse** An ascertainment bias where the genetic effect size is overestimated within the discovery dataset

## GWAS from diverse populations

The lack of diversity in GWAS has been well documented [10–13] and has spurred new efforts to increase global representation, including the Human Health and Hereditary initiative in Africa [14], the All of Us programme in the USA [15] and new initiatives such as the ‘Latin American Alliance for Genomic Diversity’ (International Common Disease Alliance [ICDA], plenary programme 2020), the Taiwan precision medicine

initiative [16] and The Brazilian Initiative on Precision Medicine (BIPMed) [17].

**Allelic frequency differences between populations aid locus discovery** GWAS in diverse populations has facilitated the discovery of novel type 2 diabetes aetiological factors owing to their divergent allele frequency across populations. One example is the risk haplotype near *SLC16A11*, discovered in Mexicans, that has high frequency in populations from the Americas (~50%), intermediate frequency in East Asians

(~10%) and is rare or absent in populations from Europe and Africa [18]. Other examples include the rare Glu508Lys variant in *HNF1A* identified in Latinos that increases type 2 diabetes risk fivefold [19] and the East Asian Arg193His *PAX4* variant [20].

The largest analyses of type 2 diabetes in African Americans to date identified novel African American signals at *HLA-B* and *INS-IGF2* loci [21]. GWAS of cardiometabolic traits including African participants are still few [22–26] but further highlight a type 2 diabetes risk variant at *ZRANB3*, which is monomorphic elsewhere [24], and new African signals at *TCF7L2* (rs17746147) and near *AGMO* (rs73284431 [23]). A pan-African GWAS of 34 cardiometabolic traits that included 14,126 individuals identified a variant driven by the  $\alpha^{-3.7}$  thalassaemia deletion associated with HbA<sub>1c</sub> in Ugandans [25]. This deletion is more frequent in Ugandans as it confers resistance to severe malaria, which is endemic in Uganda [25].

These are important examples of population-specific signals (i.e. signals where the variant is very rare or monomorphic outside the cognate population, or signals where the effect of the variant on the trait has not been observed outside those cognate populations). Nonetheless, they can reveal population-specific disease aetiology, provide novel insights into pathophysiological pathways involved in disease and highlight novel aspects of biology not previously understood.

#### Population-specific signals may be clinically important

Population-specific signals can identify variants that have large effects in cognate populations and, hence, may have an important translational impact in those populations. For example, the *TBC1D4* nonsense variant p.Arg684ter was initially found in Inuits from Greenland [27], where it has a high prevalence (17%) and large effect size (homozygous carriers have an approximately tenfold increased risk of type 2 diabetes), but is very rare or absent elsewhere. The same variant has now been detected at high frequency (~13–16% minor allele frequency) in North American Inuit populations. Here, it was shown that unless postprandial glucose levels were tested, 32% of *TBC1D4* p.Arg684ter carriers with prediabetes (defined as fasting plasma glucose 5.6–6.9 mmol/l, 2 h 75 g OGTT plasma glucose 7.8–11.0 mmol/l and/or HbA<sub>1c</sub> 5.7–6.4% [39–46 mmol/mol]) and diabetes would remain undiagnosed [28]. In light of increasing diabetes prevalence in the Inuit [29], it has been suggested that stratifying diabetes diagnoses based on an individual's *TBC1D4* p.Arg684ter genotype, and performing OGTTs in carriers of this variant, may be appropriate in this population [28]. In addition, *TBC1D4* acts on the insulin-stimulated glucose response

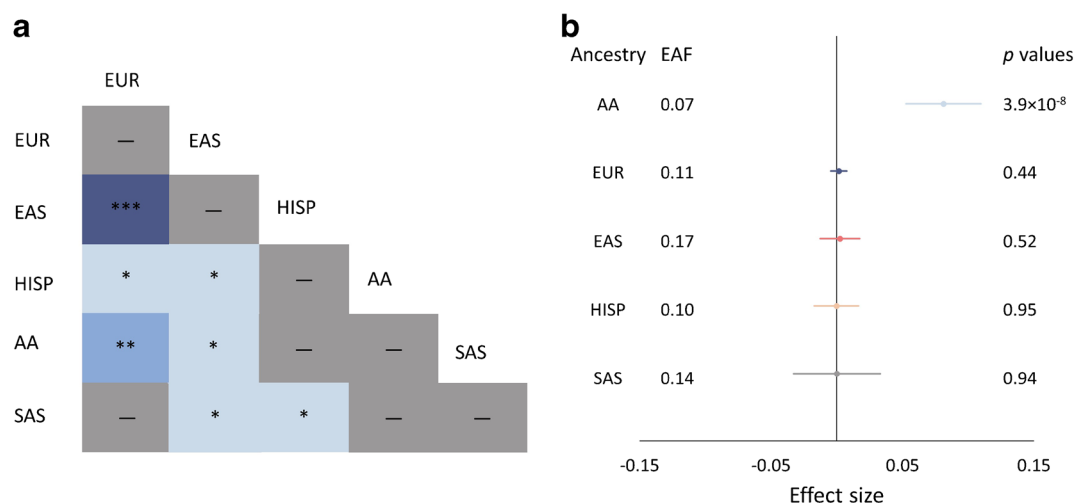
pathway so it is plausible that carriers for this variant will have improved response to insulin sensitisers, although clinical trials have yet to be performed to test this [28]. On the other hand, a recent longitudinal analysis of Inuits in Greenland suggested that homozygosity for *TBC1D4* p.Arg684ter did not significantly increase risk of incident CVD in this population [30]. Given the small number of homozygous *TBC1D4* p.Arg684ter individuals in the study ( $n = 142$ ), the possible inaccuracy in defining CVD outcomes, insufficient number of follow-up years, or other factors discussed by the authors [30], it is critical to replicate this finding. If replicated, this could suggest that diabetes associated with homozygosity for *TBC1D4* p.Arg684ter is similar to MODY due to *GCK* mutations [31], and would impact on how diabetes is managed in individuals homozygous for *TBC1D4* p.Arg684ter. Overall, this example highlights the potential importance of capturing population-specific signals for precision medicine approaches in diabetes diagnosis.

#### Interpreting population-specific signals can be challenging

Establishing the broader relevance and reproducibility of population-specific signals, especially those that result from sequence-based rare variant analysis, can be difficult. This is because due to founder effects, drift and selection, population isolates are enriched for alleles that may be very rare or absent elsewhere [32]. In addition, some indigenous specific variants originate from discovery sample sizes in the order of thousands rather than hundreds of thousands and large population resources for replication are not always readily available. Naturally, larger effect sizes in these population-specific signals are not uncommon, as these are the effect sizes some of these smaller discovery samples are well-powered to detect. In these scenarios, given the high multiple testing burden, the lower power and the absence of replication datasets, it can be hard to distinguish between true population-specific signals and false-positive associations.

#### Genome-wide multi-ancestry genetic analyses

Recently, efforts to jointly analyse different genetic datasets from populations of diverse ancestry have become more widespread [5, 6]. These multi-ancestry genetic analyses boost power for new locus discovery, provide the opportunity to test for widespread replication of signals across independent populations and allow exploration of the genetic architecture of phenotypes across ancestries.



**Fig. 1** Fasting glucose lead variants with evidence of effect allele heterogeneity across populations of different ancestry. **(a)** Fasting glucose-associated lead variants were tested for evidence of effect allele heterogeneity between populations. The findings from the test of effect allele heterogeneity are shown; a one-side heterogeneity test without multiple testing corrections was conducted and different shades of blue represent different  $p$  value thresholds (the darker the shade of blue, the more significant the  $p$  value).  $*p < 1 \times 10^{-4}$  to  $\leq 0.05$ ;  $**p < 1 \times 10^{-6}$  to  $\leq 1 \times 10^{-4}$ ;  $***p \leq 1 \times 10^{-6}$  (dash [—] represents  $p > 0.05$ ). **(b)** Forest plot

showing the effect allele frequency, effect size, 95% CIs and  $p$  value for rs61909476, the lead variant associated with fasting glucose in participants of African American ancestry. The same variant shows no evidence of association with fasting glucose in the other ancestries included in the analyses. AA, African American ancestry; EAF, effect allele frequency; EAS, East Asian ancestry; EUR, European ancestry; HISP, Hispanic ancestry; SAS, South Asian ancestry. Adapted from [6]. This figure is available as part of a [downloadable slideset](#)

**Portability of signals across populations** Evidence to date suggests that most common variants associated with type 2 diabetes or continuous glycaemic traits are shared and have broadly equivalent effects across ancestries [6, 33]. However, the Population Architecture using Genomics and Epidemiology (PAGE) Consortium showed significant effect size attenuation at established loci in non-Europeans. As effect sizes were differentially attenuated between ancestries (by ~56% in African Americans and ~24% in Hispanics/Latinos), this suggested the attenuation was not just due to ‘winner’s curse’ [34]. Recently, in a large multi-ancestry meta-analysis, we also found evidence of effect size heterogeneity between populations, in approximately 20% of loci associated with glycaemic traits [6]. For example, we detected significant evidence of effect allele heterogeneity at fasting glucose lead variants between European and East Asian ancestry participants (Fig. 1a). In addition, we found novel loci that had broadly similar allele frequency but with significant effect size differences across ancestries and evidence of association at single ancestries. The variant rs61909476, near *ETSI*, is associated with fasting glucose in African American individuals but not in those from any of the other ancestries, despite broadly similar allele frequency across ancestries (Fig. 1b) [6]. Effect size differences between ancestries can occur because the variant is tagging a causal variant more strongly in one ancestry or because there are population-specific genetic epistatic effects (i.e. genotype-by-genotype interactions) or genotype-by-environment interactions.

**Benefits and challenges of multi-ancestry studies** Multi-ancestry approaches have improved global representation, vastly increased the total sample size of type 2 diabetes and related quantitative trait genetic studies, and have yielded additional associated loci that have effects across populations from multiple ancestries. Often the variant identified from combined multi-ancestry analysis does not meet stringent genome-wide significance thresholds in individual contributing ancestries but there is still evidence that it captures a proportion of the heritability of that trait in that ancestry (Fig. 2) [2, 4–6]. Specifically, a recent study by the Meta-Analysis of Glucose and Insulin-related traits Consortium (MAGIC), which included 30% non-European ancestry participants, showed that including lead variants identified from the meta-analysis across ancestries in a genetic score captured more of the trait variance than the more limited set of variants that met stringent genome-wide significant thresholds in that population (Fig. 2) [6].

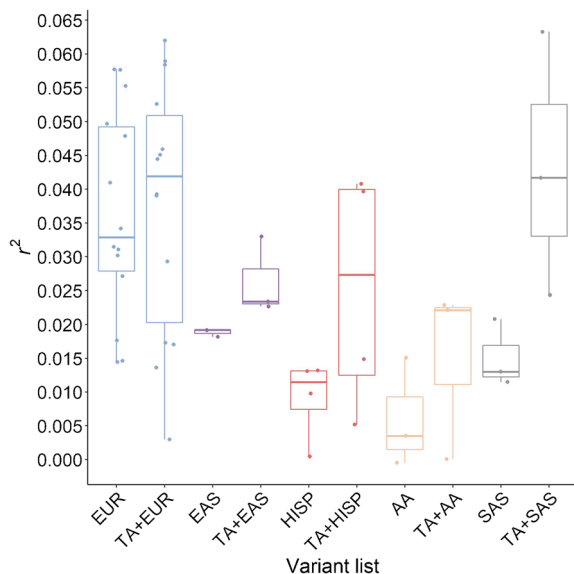
Challenges of genetic meta-analyses across ancestries relate to differences in linkage disequilibrium (LD) between populations of diverse ancestry. In this setting, clumping of variants into loci cannot be done by LD but rather by grouping together variants based on a predefined distance to the lead variant. One of the challenges of combining diverse population data is that the lead variant can vary between ancestries. Interpreting this can be difficult as it could result from random fluctuation (e.g. slightly different samples with good quality

genotype data at each variant) or different tagging of the underlying causal variant, or it could reflect allelic heterogeneity.

**Fine-mapping** The high degree of LD in European populations is both an advantage and disadvantage when conducting GWAS. High LD between variants is beneficial when conducting locus discovery as many correlated variants can point to a strong association signal. However, a consequence of this is that many variants are indistinguishable from one another in terms of their association with a disease/trait and it can therefore be difficult to establish which is the variant(s) driving the association (causal variant[s]). Fine-mapping is improved through increasing sample size so that LD can be ‘broken’ and smaller sets of variants can be identified. An approach which has gained interest is the use of populations of diverse ancestry to refine association signals [35]. Given that the LD structure differs between populations of different

ancestry, this can be leveraged to refine association signals and reduce the number of variants that need to be considered as possibly causal. This has facilitated researchers in resolving association signals to identify a smaller number of likely causal variants that can more reasonably be experimentally tested for functional effects [5, 6, 35–37].

However, fine-mapping across ancestries assumes no allelic heterogeneity at the locus being fine-mapped and assumes the causal variant(s) is shared across all populations used. Consequently, where there is true allelic heterogeneity fine-mapping across ancestries may fail. In addition, there may be technical challenges as many methods relying on summary statistics require that all variants used in the fine-mapping step have data from broadly similar sample sizes, otherwise they may identify false-positive causal variants. Moreover, removal of variants due to quality control issues could inadvertently remove the true causal variant. Nonetheless, fine-mapping methods may still identify a set of variants with high probability of being causal, which may lead researchers to follow an incorrect set of variants in downstream analyses. Comparing results from the fine-mapping to the original meta-analysis within and across ancestries is therefore key to ensure that the lead variant(s), for example, are still within the set of likely causal variants after fine-mapping. Fine-mapping across ancestries can also be challenging, as many methods are not able to account for the heterogeneity in LD across ancestries. An important challenge is that different fine-mapping methods will yield different results so, ultimately, functional validation is required to validate causal variants. Finally, phenotype heterogeneity could also underlie some differences across populations. This is less pertinent to quantitative trait measures that are well standardised but can complicate interpretation in disease studies if cases are ascertained based on very different criteria.



**Fig. 2** Variance in fasting glucose explained by associated fasting glucose loci. The box and whisker plot shows the trait variance ( $r^2$ ) explained when using a genetic score with variants that are associated with fasting glucose at genome-wide significant thresholds in each individual ancestry (European [EUR;], East Asian [EAS;], Hispanic [HISP;], African American [AA;] or Southeast Asian [SAS;]), or when using a combination of lead variants identified by meta-analysis across all participating ancestries (trans-ancestry [TA]) and individual ancestry genome-wide associated variants (TA+EUR, TA+EAS, TA+HISP, TA+AA and TA+SAS). Variance in fasting glucose explained by each of the variant lists in each individual ancestry is shown in blue (EUR), purple (EAS), red (HISP), orange (AA) and grey (SAS). The line within each box represents the median, and the top and bottom of the box represent the first and third quartile, respectively. The whiskers indicate the maximum and minimum values. Data points represent the variance explained in individual cohorts used in the analysis. Figure from [6]. This figure is available as part of a [downloadable slideset](#)

## The importance of conducting studies across multiple ancestries for precision medicine

In contrast to existing approaches to medicine that have been described as ‘one size fits all’, precision medicine proposes to take into account individual differences in genetic makeup, environment and lifestyle when considering disease presentation, diagnoses, treatment and prevention [38].

Historically, there has been limited representation of individuals of diverse ancestry in biobanks, in clinical trials [39] and, as discussed earlier, in genetic studies. Lack of representation in studies means that diagnostic thresholds, treatment regimens and prediction models do not consider genetic



differences between ancestries. This means that most of the health and economic benefit from genetics-driven approaches to medicine will inequitably benefit higher income countries (and within those, individuals of European descent), increasing health disparities between diverse populations [40].

**Impact of individual variants on diagnosis, treatment response and adverse drug reactions** In addition to the *TBC1D4* nonsense mutation, discussed above, which may have important implications for diabetes diagnosis in Inuit populations, there are other examples of ancestry-differentiated variants with impact on type 2 diabetes diagnosis and treatment.

The *G6PD* Val98Met (rs1050828) variant causes glucose 6-phosphate dehydrogenase deficiency, a haemolytic anaemia that is often silent in carriers (i.e. they may not know they have the mutation). The same variant reduces HbA<sub>1c</sub> levels ( $\beta = -0.81\%$  [95% CI 0.66, 0.96] per minor allele) independently of glucose levels and potentially leads to under-diagnoses of diabetes in carriers [41, 42]. Other *G6PD* variants that lower HbA<sub>1c</sub> levels have been identified in Hispanic/Latino [43] and Asian populations [44]. In addition, carriers of the  $\alpha^{-3.7}$  thalassaemia deletion [25] and asymptomatic individuals with sickle cell trait (rs334 Glu7Val) [43, 45] all have reduced HbA<sub>1c</sub> levels, independent of glucose levels.

Some of the above variants are common (minor allele frequency >10%) in populations with endemic malaria, as they provide protection against severe malaria [46–48], and all affect the utility of HbA<sub>1c</sub> as a diagnostic test for diabetes in those populations. Because the prevalence of these variants differs between ancestral groups, ignoring genotype at these variants could exacerbate health disparities. In addition, in carriers being treated for diabetes, physicians may overestimate the degree of glucose control (as carriers will have disproportionately low HbA<sub>1c</sub> for their blood glucose levels) and therefore undertreat [44].

Beyond effects on diagnosis and treatment targets, the *G6PD* Val98Met variant is associated with significant risk of haemolysis in women treated with the antimalarial agent primaquine [49] and the US Food and Drug Administration has declared the need to consider *G6PD* status for patients prescribed certain sulfonylureas [50], highlighting the importance of knowing genotype at this site before prescribing drugs.

**The promise of genetic risk scores** Variants that associate at genome-wide significant levels with a trait or a disease can be used to construct genetic risk scores (GRSs) that explain or predict a certain proportion of the trait variance in the

population [51, 52]. The hope is that these scores may have clinical utility by facilitating identification of individuals at higher risk of disease, aiding in differential diagnoses, better targeting of treatment and therapy dosage to patients, and helping to avoid adverse drug reactions.

Early type 2 diabetes GRSs were built on a limited set of variants, explained a relatively modest fraction of phenotypic variance and were not very useful for disease prediction [53–56]. Additionally, as they were mostly built on results from large meta-analyses of European ancestry GWAS, they missed the effects of other ancestry-specific trait-associated variants, namely variants under different types of selection in populations exposed to different environments. Furthermore, the effect size estimates used were overinflated due to ‘winner’s curse’ in discovery studies [57].

However, as sample sizes increased, more variants have been detected that capture more of trait variance. In addition, when genome-wide associated variants from multi-ancestry studies are used to build GRSs, they capture a larger fraction of phenotypic variance than ancestry-specific GRSs even if the variants are not associated with the disease at genome-wide significance level in all ancestries [5, 6, 58]. This suggests that such multi-ancestry efforts may be required for GRSs to be more globally transferable.

**A type 1 diabetes GRS with clinical utility** Provision of the correct diabetes diagnosis is important, as the optimal treatment is different for type 1 diabetes, type 2 diabetes and other rare monogenic forms of diabetes. Here, a type 1 diabetes GRS (and subsequent successor) has clinical utility, improving newborn screening and supporting classification of adult incident diabetes in individuals of European ancestry [59, 60]. It also helps differentiate between type 1 diabetes and monogenic neonatal diabetes or MODY [61] and monogenic autoimmune diabetes [62]. Despite early concerns regarding the transferability of the GRS to other ancestries [59], the GRS was shown to discriminate between monogenic and type 1 diabetes in Iranian children [63]. It also discriminates between type 1 and type 2 diabetes in India, where misclassification of type 1 diabetes and type 2 diabetes is common in young adults due to the high prevalence of early-onset type 2 diabetes at lower BMI [64].

**Polygenic risk scores** Beyond genome-wide significant variants, models that include additional loci in the genome that have not reached this stringent threshold capture a larger fraction of trait variance [65–67]. These variants have been included in polygenic risk scores (PRSs), which are built on a large number of variants in the genome (in the order of thousands to several million), to improve disease prediction [68–70] (see Text box: Concerns regarding transferability of PRSs).

## Concerns regarding transferability of PRSs

Over the years, different methods and associated software packages have been developed to derive PRSs [87–90]. These require developing the PRS based on one dataset, validating it in an independent dataset, and finally applying it to the desired target population.

These methods use different algorithms to select the best combination of SNPs across the entire genome that maximise the phenotypic variance explained. Or, in other words, it selects the combination of SNPs that best captures phenotype variation in that population.

To work, these methods have to account for the LD between SNPs so they do not ‘double count’ effects of highly correlated SNPs (SNPs with highly correlated effects). The challenge is that, when using the same PRS in other populations of different ancestry, because the underlying LD between variants is different, the PRS will not perform as well in new target populations [57, 91–93]. This is because in the new target populations, the PRS no longer appropriately corrects for the different correlation between the SNPs in the score. The more divergent the discovery and target populations, the worse a PRS will perform from one to the other. Because PRSs are normally built on European ancestry training datasets, their implementation in the clinic has the potential to systematically benefit individuals of European descent over people from other ancestries [57, 93].

The differences in effect size estimates due to LD differences between training and target populations are exacerbated in PRSs (vs GRSs) because they sum over all these effects on thousands or millions of variants, therefore, aggregating population differences relative to GRSs built on a much more limited number of genome-wide significant variants that are enriched in causal variants (or good tags for causal variants) that tend to be shared, and have equivalent effects, between populations [93].

These issues with PRSs can also arise when the training and target populations are from the same broad ancestry but where the training set is derived from the aggregation of a very large collection of datasets in which there is residual substructure or population stratification [94, 95]. Population substructure occurs when there are systematic differences in allelic frequency between subgroups within the population. The recommendation here might be to favour the use of more homogeneous large biobank populations as a training dataset over the use of large GWAS meta-analyses as the training dataset. This, however, will not always be an option and it remains a concern in the development and application of PRSs in the clinic.

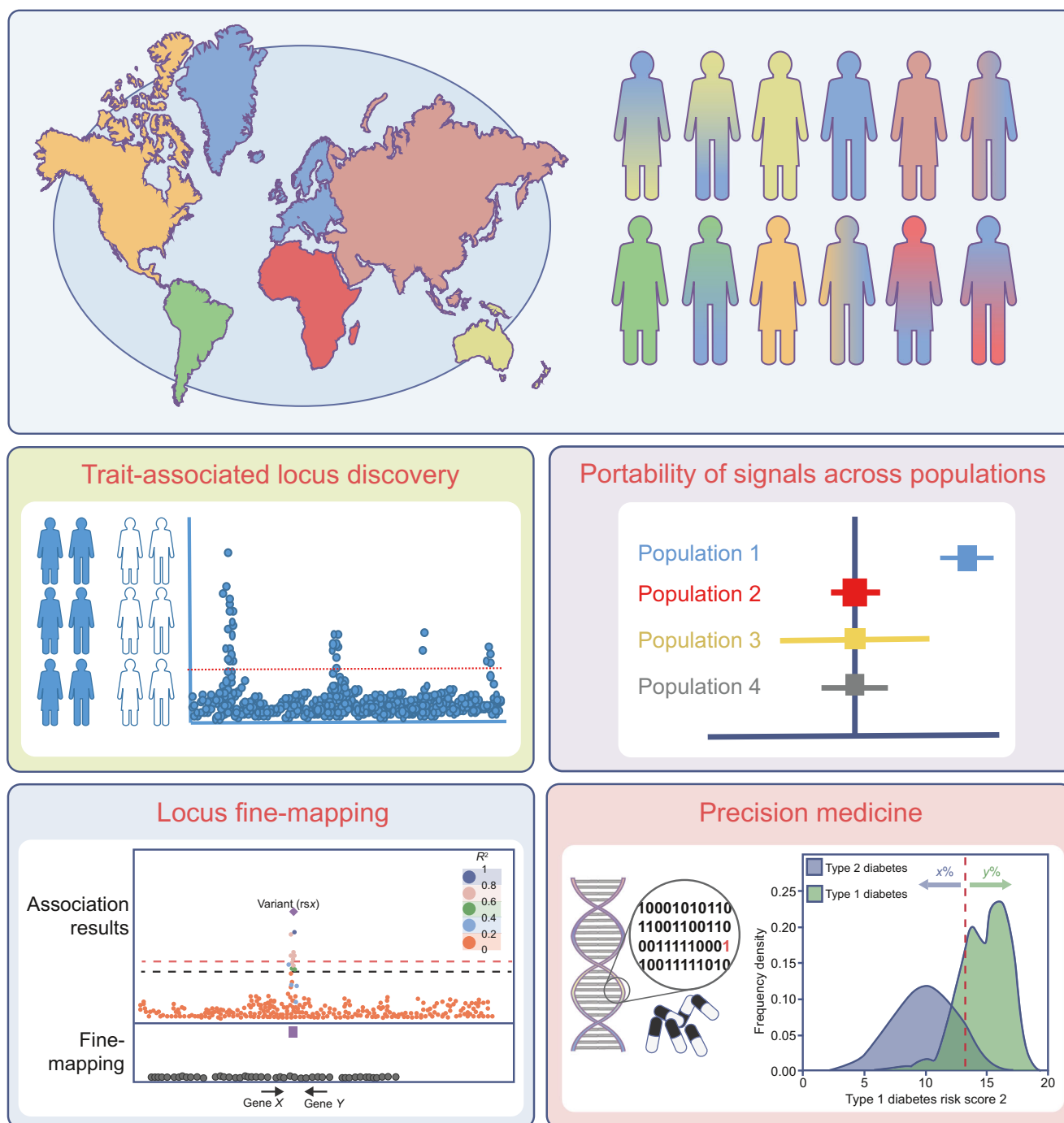
A proposed potential benefit of PRSs is their ability to identify high-risk individuals from birth before classical clinical and biomarker risk factors can be detected, thus enabling identification of a subset of the population who would most benefit from careful screening and monitoring, and from being placed on available preventative strategies or therapies [56]. They may also identify individuals at much larger risk of disease who might not display classical clinical risk factors and would, hence, be missed by current approaches [56], though this has been questioned [71].

Individuals at the top end of the distribution for type 2 diabetes PRSs have a disease risk similar to that of individuals harbouring some monogenic mutations [8, 56]. Nevertheless, concerns regarding portability of these scores across populations (see Text box: Concerns regarding transferability of PRSs) raise doubt over their current clinical utility, and provide a compelling argument for developing scores based on discovery data from diverse populations, as these more readily transfer from discovery to different target populations [57, 58]. Given these concerns, at least for type 1 and type 2

diabetes, it has been suggested by some that GRSs (especially those arising from multi-ancestry analyses) may be currently preferable, as the additional variants in the PRS do not significantly improve the performance of these scores for clinical use [71, 72].

Beyond the cross-population transferability issues that may or may not be addressed by further methodological development, questions remain regarding how PRSs predict disease risk across the lifespan [73, 74], how risk is understood and communicated by practicing clinicians to their patients and, more generally, how to incorporate their use into routine clinical practice [75]. Indeed, the debate rages on as to whether these scores will provide broad clinical utility beyond a few examples [68, 71, 76, 77].

**Partitioned genetic scores** In addition to the use of GRSs and PRSs for disease classification and prediction, the development of partitioned genetic scores corresponding to variants predicted to affect disease through different physiological pathways has gained interest as a means to acquire insight into



**Fig. 3** Population diversity in genetic studies of type 2 diabetes and related glycaemic traits. The diagram shows a pictorial representation of the world, with its populations and their admixture represented by the shaded people; the different colours represent differences in ancestral admixture in different individuals. The main areas that benefit from increasing population diversity in genetic studies are shown; these

include: trait-associated locus discovery; portability of signals across populations; improving fine-mapping resolution; and development of a more equitable precision medicine approach (e.g. through development of GRSs or PRSs based on multi-ancestral population data). This figure is available as part of a [downloadable slideset](#)

disease heterogeneity [78, 79]. These partitioned scores may be able to identify subsets of individuals with type 2 diabetes having different risks of complications [79]. Possible clinical utility could additionally result from patient stratification for correct treatment and therapy dosage according to the major

pathway predicted to be affected in the subset of patients, and for the identification of participants for clinical trials [56]. However, whether patient stratification for treatment will follow the success seen in monogenic diabetes remains a big question in the field [72].



## Conclusions

Over the last few years, diabetes and glycaemic trait GWAS have included data with broader genetic diversity. This has led to novel locus discovery, improved understanding of the genetic architecture of diabetes and related glycaemic traits across ancestries, improved fine-mapping resolution and resulted in the development of GRSs that better capture disease risk across populations (Fig. 3). Nevertheless, efforts to increase global representation in genetic studies need to be intensified to fully capture the aetiology of type 2 diabetes and associated traits across the world, specifically in under-represented populations, wherein the rise in diabetes prevalence is predicted to be especially notable in the forthcoming years. There is a need to increase representation of different ancestries in regulatory annotation efforts (e.g. generation of expression quantitative trait [eQTL] data), to enable ancestry-specific effects to be interpreted within local context. These annotations have been instrumental in pinpointing causal genes at GWAS loci [8] and are key in the journey from genetic association to improved mechanistic insight.

The opportunities afforded by increasing diversity in genetic studies of type 2 diabetes and related glycaemic traits are undisputed. However, it is important to stress that the human population is a continuum with no discrete boundaries between groups, whether these are defined on the basis of self-reported ethnicity or on the basis of genetically defined ancestry. It is critical therefore, that we move away from describing ancestry based on large continental labels and acknowledge the finer-grained population-level genetic diversity that reflects population history, migration and admixture. Though there are practical reasons for grouping individuals into clusters, in the end we are all admixed with different degrees of contribution from various ancestral groups.

Though significant progress has been made, there remain methodological challenges relating to allelic, phenotypic and environmental heterogeneity. Most importantly, there are significant ethical, societal and cultural challenges still to overcome. Given historical malpractices [80, 81], some communities have naturally become disengaged and suspicious of genetic and genomic efforts. Going forward, engaging with global and indigenous populations needs to be done sensitively and be respectful of local cultures. Ownership of the research agenda and leadership has to be held by those within those communities [82]. An example is the H3A initiative, which set out to empower African researchers to lead and take centre stage in genomic research [83–85]. A considered balance needs to be achieved between the desire to rapidly, publicly, share data globally for the advancement of science and the need to consider critical aspects of indigenous governance policies for self-determination with respect to genomics issues [81]. In addition, equity of access, and ability to use the samples collected and analyse the data generated are important to help level out the playing field [86].

It must also be recognised that for global collaborations between high-income and low- and middle-income countries to be effective, one must take the view of the importance of long-term deliverables rather than focus exclusively on short-term gains. Investment must be made in infrastructure, in building local research capacity and leadership, and in creating opportunities for ‘brain gain’. New initiatives that perhaps focus on bringing experts in from outside for periods of time to conduct research locally, collaborate, train and build local capacity instead of taking local researchers or samples out may avoid so-called ‘helicopter’ or ‘parachute’ science [80, 81]. However, it must be recognised that progress will take time and will need to leverage outside funding to generate investment from local governments. In sum, the road ahead may be long and arduous but it will surely lead us to a better world.

**Supplementary Information** The online version contains a slideset of the figures for download available at <https://doi.org/10.1007/s00125-021-05575-4>.

**Acknowledgements** The author would like to thank J. Chen (University of Exeter Medical School, Exeter, UK) for help with the figures.

**Funding** IB acknowledges support from an ‘Expanding excellence in England’ award from Research England.

**Author’s relationships and activities** The author declares that there are no relationships or activities that might bias, or be perceived to bias, this work.

**Contribution statement** The author was the sole contributor to this manuscript.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. International Diabetes Federation (2019) IDF Diabetes Atlas, 9th edn. International Diabetes Federation, Brussels
2. Spracklen CN, Horikoshi M, Kim YJ et al (2020) Identification of type 2 diabetes loci in 433,540 East Asian individuals. *Nature* 582(7811):240–245. <https://doi.org/10.1038/s41586-020-2263-3>
3. Ishigaki K, Akiyama M, Kanai M et al (2020) Large-scale genome-wide association study in a Japanese population identifies novel

- susceptibility loci across different diseases. *Nat Genet* 52(7):669–679. <https://doi.org/10.1038/s41588-020-0640-3>
4. Vujkovic M, Keaton JM, Lynch JA et al (2020) Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat Genet* 52(7):680–691. <https://doi.org/10.1038/s41588-020-0637-y>
  5. Mahajan A, Spracklen CN, Zhang W et al. (2020) Trans-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *medRxiv*: 2020.2009.2022.20198937 (Preprint). 23 September 2020. Available from <https://doi.org/10.1101/2020.09.22.20198937>. Accessed 1 Dec 2020
  6. Chen J, Spracklen CN, Marenne G et al (2021) The trans-ancestral genomic architecture of glycemic traits. *Nat Genet* 53(6):840–860. <https://doi.org/10.1038/s41588-021-00852-9>
  7. Scott RA, Lagou V, Welch RP et al (2012) Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat Genet* 44(9):991–1005. <https://doi.org/10.1038/ng.2385>
  8. Mahajan A, Taliun D, Thurner M et al (2018) Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat Genet* 50(11):1505–1513. <https://doi.org/10.1038/s41588-018-0241-6>
  9. Batai K, Hooker S, Kittles RA (2020) Leveraging genetic ancestry to study health disparities. *Am J Phys Anthropol* 175(2):363–375. <https://doi.org/10.1002/ajpa.24144>
  10. Need AC, Goldstein DB (2009) Next generation disparities in human genomics: concerns and remedies. *Trends Genet* 25(11):489–494. <https://doi.org/10.1016/j.tig.2009.09.012>
  11. Popejoy AB, Fullerton SM (2016) Genomics is failing on diversity. *Nature* 538(7624):161–164. <https://doi.org/10.1038/538161a>
  12. Morales J, Welter D, Bowler EH et al (2018) A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. *Genome Biol* 19(1):21. <https://doi.org/10.1186/s13059-018-1396-2>
  13. Mills MC, Rahal C (2020) The GWAS Diversity Monitor tracks diversity by disease in real time. *Nat Genet* 52(3):242–243. <https://doi.org/10.1038/s41588-020-0580-y>
  14. The H3Africa Consortium, Rotimi C, Abayomi A et al (2014) Research capacity. Enabling the genomic revolution in Africa. *Science* 344(6190):1346–1348. <https://doi.org/10.1126/science.1251546>
  15. Collins FS, Varmus H (2015) A new initiative on precision medicine. *N Engl J Med* 372(9):793–795. <https://doi.org/10.1056/NEJMp1500523>
  16. Taiwan Precision Medicine Initiative (2020). Taiwan Precision Medicine Initiative. Available from <https://tpmi.ibms.sinica.edu.tw/www/en/>. Accessed 1 Dec 2020
  17. Rocha CS, Secolin R, Rodrigues MR, Carvalho BS, Lopes-Cendes I (2020) The Brazilian Initiative on Precision Medicine (BIPMed): fostering genomic data-sharing of underrepresented populations. *NPJ Genom Med* 5:42. <https://doi.org/10.1038/s41525-020-00149-6>
  18. The SIGMA Type 2 Diabetes Consortium (2014) Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature* 506(7486):97–101. <https://doi.org/10.1038/nature12828>
  19. The SIGMA Type 2 Diabetes Consortium (2014) Association of a low-frequency variant in HNF1A with type 2 diabetes in a Latino population. *JAMA* 311(22):2305–2314. <https://doi.org/10.1001/jama.2014.6511>
  20. Fuchsberger C, Flannick J, Teslovich TM et al (2016) The genetic architecture of type 2 diabetes. *Nature* 536(7614):41–47. <https://doi.org/10.1038/nature18642>
  21. Ng MC, Shriner D, Chen BH et al (2014) Meta-analysis of genome-wide association studies in African Americans provides insights into the genetic architecture of type 2 diabetes. *PLoS Genet* 10(8):e1004517. <https://doi.org/10.1371/journal.pgen.1004517>
  22. Adeyemo AA, Tekola-Ayele F, Doumatey AP et al (2015) Evaluation of Genome Wide Association Study Associated Type 2 Diabetes Susceptibility Loci in Sub Saharan Africans. *Front Genet* 6:335. <https://doi.org/10.3389/fgene.2015.00335>
  23. Chen J, Sun M, Adeyemo A et al (2019) Genome-wide association study of type 2 diabetes in Africa. *Diabetologia* 62(7):1204–1211. <https://doi.org/10.1007/s00125-019-4880-7>
  24. Adeyemo AA, Zaghoul NA, Chen G et al (2019) ZRANB3 is an African-specific type 2 diabetes locus associated with beta-cell mass and insulin response. *Nat Commun* 10(1):3195. <https://doi.org/10.1038/s41467-019-10967-7>
  25. Gurdasani D, Carstensen T, Fatumo S et al (2019) Uganda Genome Resource Enables Insights into Population History and Genomic Discovery in Africa. *Cell* 179(4):984–1002 e1036. <https://doi.org/10.1016/j.cell.2019.10.004>
  26. Liu C, Chen G, Bentley AR et al (2019) Genome-wide association study for proliferative diabetic retinopathy in Africans. *NPJ Genom Med* 4:20. <https://doi.org/10.1038/s41525-019-0094-7>
  27. Moltke I, Grarup N, Jorgensen ME et al (2014) A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. *Nature* 512(7513):190–193. <https://doi.org/10.1038/nature13425>
  28. Manousaki D, Kent JW Jr, Haack K et al (2016) Toward Precision Medicine: TBC1D4 Disruption Is Common Among the Inuit and Leads to Underdiagnosis of Type 2 Diabetes. *Diabetes Care* 39(11):1889–1895. <https://doi.org/10.2337/dc16-0769>
  29. Jorgensen ME, Bjeregaard P, Borch-Johnsen K (2002) Diabetes and impaired glucose tolerance among the inuit population of Greenland. *Diabetes Care* 25(10):1766–1771. <https://doi.org/10.2337/diacare.25.10.1766>
  30. Overvad M, Diaz LJ, Bjerregaard P et al (2020) The effect of diabetes and the common diabetogenic TBC1D4 p.Arg684Ter variant on cardiovascular risk in Inuit in Greenland. *Sci Rep* 10(1):22081. <https://doi.org/10.1038/s41598-020-79132-1>
  31. Steele AM, Shields BM, Wensley KJ, Colclough K, Ellard S, Hattersley AT (2014) Prevalence of vascular complications among patients with glucokinase mutations and prolonged, mild hyperglycemia. *JAMA* 311(3):279–286. <https://doi.org/10.1001/jama.2013.283980>
  32. Hatzikotoulas K, Gilly A, Zeggini E (2014) Using population isolates in genetic association studies. *Brief Funct Genomics* 13(5):371–377. <https://doi.org/10.1093/bfpg/elu022>
  33. Barroso I, McCarthy MI (2019) The Genetic Basis of Metabolic Disease. *Cell* 177(1):146–161. <https://doi.org/10.1016/j.cell.2019.02.024>
  34. Wojcik GL, Graff M, Nishimura KK et al (2019) Genetic analyses of diverse populations improves discovery for complex traits. *Nature* 570(7762):514–518. <https://doi.org/10.1038/s41586-019-1310-4>
  35. Bien SA, Pankow JS, Haessler J et al (2017) Transethnic insight into the genetics of glycaemic traits: fine-mapping results from the Population Architecture using Genomics and Epidemiology (PAGE) consortium. *Diabetologia* 60(12):2384–2398. <https://doi.org/10.1007/s00125-017-4405-1>
  36. Helgason A, Palsson S, Thorleifsson G et al (2007) Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. *Nat Genet* 39(2):218–225. <https://doi.org/10.1038/ng1960>
  37. DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, Mexican American Type 2

- Diabetes (MAT2D) Consortium, Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES) Consortium (2014) Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* 46(3): 234–244. <https://doi.org/10.1038/ng.2897>
38. Feero WG (2017) Introducing “Genomics and Precision Health”. *JAMA* 317(18):1842–1843. <https://doi.org/10.1001/jama.2016.20625>
  39. Chen MS Jr, Lara PN, Dang JH, Paterniti DA, Kelly K (2014) Twenty years post-NIH Revitalization Act: enhancing minority participation in clinical trials (EMPaCT): laying the groundwork for improving minority clinical trial accrual: renewing the case for enhancing minority participation in cancer clinical trials. *Cancer* 120(Suppl 7):1091–1096. <https://doi.org/10.1002/cncr.28575>
  40. Smith CE, Fullerton SM, Dookeran KA et al (2016) Using Genetic Technologies To Reduce, Rather Than Widen, Health Disparities. *Health Aff* 35(8):1367–1373. <https://doi.org/10.1377/hlthaff.2015.1476>
  41. Wheeler E, Leong A, Liu CT et al (2017) Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. *PLoS Med* 14(9):e1002383. <https://doi.org/10.1371/journal.pmed.1002383>
  42. Sarnowski C, Leong A, Raffield LM et al (2019) Impact of Rare and Common Genetic Variants on Diabetes Diagnosis by Hemoglobin A1c in Multi-Ancestry Cohorts: The Trans-Omics for Precision Medicine Program. *Am J Hum Genet* 105(4):706–718. <https://doi.org/10.1016/j.ajhg.2019.08.010>
  43. Moon JY, Louie TL, Jain D et al (2019) A Genome-Wide Association Study Identifies Blood Disorder-Related Variants Influencing Hemoglobin A1c With Implications for Glycemic Status in U.S. Hispanics/Latinos. *Diabetes Care* 42(9):1784–1791. <https://doi.org/10.2337/dc19-0168>
  44. Leong A, Lim VJY, Wang C et al (2020) Association of G6PD variants with hemoglobin A1c and impact on diabetes diagnosis in East Asian individuals. *BMJ Open Diabetes Res Care* 8(1):e001091. <https://doi.org/10.1136/bmjdc-2019-001091>
  45. Lacy ME, Wellenius GA, Sumner AE et al (2017) Association of Sick Cell Trait With Hemoglobin A1c in African Americans. *JAMA* 317(5):507–515. <https://doi.org/10.1001/jama.2016.21035>
  46. Mockenhaupt FP, Ehrhardt S, Gellert S et al (2004)  $\alpha^+$ -Thalassemia protects African children from severe malaria. *Blood* 104(7):2003–2006. <https://doi.org/10.1182/blood-2003-11-4090>
  47. Aidoo M, Terlouw DJ, Kolczak MS et al (2002) Protective effects of the sickle cell gene against malaria morbidity and mortality. *Lancet* 359(9314):1311–1312. [https://doi.org/10.1016/S0140-6736\(02\)08273-9](https://doi.org/10.1016/S0140-6736(02)08273-9)
  48. Luzzatto L, Arese P (2018) Favism and Glucose-6-Phosphate Dehydrogenase Deficiency. *N Engl J Med* 378(1):60–71. <https://doi.org/10.1056/nejmra1708111>
  49. Chu CS, Bancone G, Moore KA et al (2017) Haemolysis in G6PD Heterozygous Females Treated with Primaquine for Plasmodium vivax Malaria: A Nested Cohort in a Trial of Radical Curative Regimens. *PLoS Med* 14(2):e1002224. <https://doi.org/10.1371/journal.pmed.1002224>
  50. FDA (2020) Table of Pharmacogenomic Biomarkers in Drug Labeling. Available from [www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm083378.htm](http://www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm083378.htm). Accessed 1 Dec 2020
  51. Janssens AC, Aulchenko YS, Elefante S, Borsboom GJ, Steyerberg EW, van Duijn CM (2006) Predictive testing for complex diseases using multiple genes: fact or fiction? *Genet Med* 8(7):395–400. <https://doi.org/10.1097/01.gim.0000229689.18263.f4>
  52. Wray NR, Goddard ME, Visscher PM (2007) Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res* 17(10):1520–1528. <https://doi.org/10.1101/gr.6665407>
  53. Lyssenko V, Jonsson A, Almgren P et al (2008) Clinical risk factors, DNA variants, and the development of type 2 diabetes. *N Engl J Med* 359(21):2220–2232. <https://doi.org/10.1056/NEJMoa0801869>
  54. Meigs JB, Shrader P, Sullivan LM et al (2008) Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N Engl J Med* 359(21):2208–2219. <https://doi.org/10.1056/NEJMoa0804742>
  55. Lango H, U.K. Type 2 Diabetes Genetics Consortium, Palmer CN et al (2008) Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. *Diabetes* 57(11):3129–3135. <https://doi.org/10.2337/db08-0504>
  56. Khera AV, Chaffin M, Aragam KG et al (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* 50(9):1219–1224. <https://doi.org/10.1038/s41588-018-0183-z>
  57. Martin AR, Gignoux CR, Walters RK et al (2017) Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am J Hum Genet* 100(4):635–649. <https://doi.org/10.1016/j.ajhg.2017.03.004>
  58. Koyama S, Ito K, Terao C et al (2020) Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic risk loci for coronary artery disease. *Nat Genet* 52(11):1169–1177. <https://doi.org/10.1038/s41588-020-0705-3>
  59. Perry DJ, Wasserfall CH, Oram RA et al (2018) Application of a Genetic Risk Score to Racially Diverse Type 1 Diabetes Populations Demonstrates the Need for Diversity in Risk-Modeling. *Sci Rep* 8(1):4529. <https://doi.org/10.1038/s41598-018-22574-5>
  60. Sharp SA, Rich SS, Wood AR et al (2019) Development and Standardization of an Improved Type 1 Diabetes Genetic Risk Score for Use in Newborn Screening and Incident Diagnosis. *Diabetes Care* 42(2):200–207. <https://doi.org/10.2337/dc18-1785>
  61. Patel KA, Oram RA, Flanagan SE et al (2016) Type 1 Diabetes Genetic Risk Score: A Novel Tool to Discriminate Monogenic and Type 1 Diabetes. *Diabetes* 65(7):2094–2099. <https://doi.org/10.2337/db15-1690>
  62. Johnson MB, Patel KA, De Franco E et al (2018) A type 1 diabetes genetic risk score can discriminate monogenic autoimmunity with diabetes from early-onset clustering of polygenic autoimmunity with diabetes. *Diabetologia* 61(4):862–869. <https://doi.org/10.1007/s00125-018-4551-0>
  63. Yaghootkar H, Abbasi F, Ghaemi N et al (2019) Type 1 diabetes genetic risk score discriminates between monogenic and Type 1 diabetes in children diagnosed at the age of <5 years in the Iranian population. *Diabet Med* 36(12):1694–1702. <https://doi.org/10.1111/dme.14071>
  64. Harrison JW, Tallapragada DSP, Baptist A et al (2020) Type 1 diabetes genetic risk score is discriminative of diabetes in non-Europeans: evidence from a study in India. *Sci Rep* 10(1):9450. <https://doi.org/10.1038/s41598-020-65317-1>
  65. Yang J, Benyamin B, McEvoy BP et al (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 42(7):565–569. <https://doi.org/10.1038/ng.608>
  66. Lee SH, Wray NR, Goddard ME, Visscher PM (2011) Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet* 88(3):294–305. <https://doi.org/10.1016/j.ajhg.2011.02.002>
  67. Yang J, Bakshi A, Zhu Z et al (2015) Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* 47(10):1114–1120. <https://doi.org/10.1038/ng.3390>



68. Torkamani A, Wineinger NE, Topol EJ (2018) The personal and clinical utility of polygenic risk scores. *Nat Rev Genet* 19(9):581–590. <https://doi.org/10.1038/s41576-018-0018-x>
69. Chatterjee N, Shi J, Garcia-Closas M (2016) Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet* 17(7):392–406. <https://doi.org/10.1038/nrg.2016.27>
70. Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park JH (2013) Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat Genet* 45(4):400–405, 405e401–403. <https://doi.org/10.1038/ng.2579>
71. Janssens A, Joyner MJ (2019) Polygenic Risk Scores That Predict Common Diseases Using Millions of Single Nucleotide Polymorphisms: Is More, Better? *Clin Chem* 65(5):609–611. <https://doi.org/10.1373/clinchem.2018.296103>
72. Udler MS, McCarthy MI, Florez JC, Mahajan A (2019) Genetic Risk Scores for Diabetes Diagnosis and Precision Medicine. *Endocr Rev* 40(6):1500–1520. <https://doi.org/10.1210/er.2019-00088>
73. Yang Q, Flanders WD, Moonesinghe R, Ioannidis JP, Guessous I, Khoury MJ (2009) Using lifetime risk estimates in personal genomic profiles: estimation of uncertainty. *Am J Hum Genet* 85(6):786–800. <https://doi.org/10.1016/j.ajhg.2009.10.017>
74. Desikan RS, Fan CC, Wang Y et al (2017) Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score. *PLoS Med* 14(3):e1002258. <https://doi.org/10.1371/journal.pmed.1002258>
75. Schork AJ, Schork MA, Schork NJ (2018) Genetic risks and clinical rewards. *Nat Genet* 50(9):1210–1211. <https://doi.org/10.1038/s41588-018-0213-x>
76. Nature Medicine (2018) GWAS to the people. *Nat Med* 24(10):1483. <https://doi.org/10.1038/s41591-018-0231-3>
77. Curtis D (2019) Clinical relevance of genome-wide polygenic score may be less than claimed. *Ann Hum Genet* 83(4):274–277. <https://doi.org/10.1111/ahg.12302>
78. Mahajan A, Wessel J, Willems SM et al (2018) Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat Genet* 50(4):559–571. <https://doi.org/10.1038/s41588-018-0084-1>
79. Udler MS, Kim J, von Grotthuss M et al (2018) Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. *PLoS Med* 15(9):e1002654. <https://doi.org/10.1371/journal.pmed.1002654>
80. Heymann DL, Liu J, Lillywhite L (2016) Partnerships, Not Parachutists, for Zika Research. *N Engl J Med* 374(16):1504–1505. <https://doi.org/10.1056/NEJMp1602278>
81. Garrison NA, Hudson M, Ballantyne LL et al (2019) Genomic Research Through an Indigenous Lens: Understanding the Expectations. *Annu Rev Genomics Hum Genet* 20:495–517. <https://doi.org/10.1146/annurev-genom-083118-015434>
82. Callaway E (2017) South Africa's San people issue ethics code to scientists. *Nature* 543(7646):475–476. <https://doi.org/10.1038/543475a>
83. Bentley AR, Callier S, Rotimi C (2019) The Emergence of Genomic Research in Africa and New Frameworks for Equity in Biomedical Research. *Ethn Dis* 29(Suppl 1):179–186. <https://doi.org/10.18865/ed.29.S1.179>
84. Yakubu A, Tindana P, Matimba A et al (2018) Model framework for governance of genomic research and biobanking in Africa - a content description. *AAS Open Res* 1:13. <https://doi.org/10.12688/aasopenres.12844.2>
85. Baichoo S, Souilmi Y, Panji S et al (2018) Developing reproducible bioinformatics analysis workflows for heterogeneous computing environments to support African genomics. *BMC Bioinformatics* 19(1):457. <https://doi.org/10.1186/s12859-018-2446-1>
86. Munung NS, Mayosi BM, de Vries J (2017) Equity in international health research collaborations in Africa: Perceptions and expectations of African researchers. *PLoS One* 12(10):e0186237. <https://doi.org/10.1371/journal.pone.0186237>
87. Euesden J, Lewis CM, O'Reilly PF (2015) PRSice: Polygenic Risk Score software. *Bioinformatics* 31(9):1466–1468. <https://doi.org/10.1093/bioinformatics/btu848>
88. Vilhjalmsdottir BJ, Yang J, Finucane HK et al (2015) Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet* 97(4):576–592. <https://doi.org/10.1016/j.ajhg.2015.09.001>
89. Shi J, Park JH, Duan J et al (2016) Winner's Curse Correction and Variable Thresholding Improve Performance of Polygenic Risk Modeling Based on Genome-Wide Association Study Summary-Level Data. *PLoS Genet* 12(12):e1006493. <https://doi.org/10.1371/journal.pgen.1006493>
90. Ge T, Chen CY, Ni Y, Feng YA, Smoller JW (2019) Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun* 10(1):1776. <https://doi.org/10.1038/s41467-019-09718-5>
91. Martin AR, Gignoux CR, Walters RK et al (2020) Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am J Hum Genet* 107(4):788–789 (Correction). <https://doi.org/10.1016/j.ajhg.2020.08.020>
92. Kim MS, Patel KP, Teng AK, Berens AJ, Lachance J (2018) Genetic disease risks can be misestimated across global populations. *Genome Biol* 19(1):179. <https://doi.org/10.1186/s13059-018-1561-7>
93. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ (2019) Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* 51(4):584–591. <https://doi.org/10.1038/s41588-019-0379-x>
94. Sohail M, Maier RM, Ganna A et al (2019) Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *Elife* 8:e39702. <https://doi.org/10.7554/eLife.39702>
95. Berg JJ, Harpak A, Sinnott-Armstrong N et al (2019) Reduced signal for polygenic adaptation of height in UK Biobank. *Elife* 8:e39725. <https://doi.org/10.7554/eLife.39725>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.